The M4 competition: Bigger. Stronger. Better.

Fotios Petropoulos[1] and Spyros Makridakis[2]

[1]School of Management, University of Bath, UK
[2]Institute for the Future (IFF), University of Nicosia, Cyprus

Another forecasting competition?

Over the last 40 years, we have witnessed significant theoretical and modelling advances in forecasting. Such advances are usually built under an array of implicit or explicit assumptions, including ones about the data generation processes. However, real data does not always follow theoretical data generation processes. Even if a true model existed for each time series, its formulation would be almost impossible due to a large number of direct and indirect factors (and their uncertainties) affecting it.

The Makridakis competitions have long been the standard in empirical forecasting research. The impact of these competitions is threefold:

- First, the Makridakis competitions inspired and motivated forecasting researchers and practitioners to develop and test their models and approaches to forecasting.
- Second, they provided a distinct, publicly available data sets. For many forecasting researchers, including the first author of this editorial, these data sets have been a "playground" where new ideas are tested, and so he celebrated when the 13.85% sMAPE of the Theta method on the M3 monthly subset was outperformed in the M4 Competition.
- Third, the M competitions bridge the gap between theory and practice.

This impact is reflected in the thousands of citations that they have received until now. The article which describes the results of the original M competition (Makridakis et al., 1982) has been cited 1371 times according to Google Scholar (as of May 11, 2019), making it the most cited article of the *Journal of Forecasting*. The M3 competition article (Makridakis and Hibon, 2000) has 1334 citations, which puts it in 5th place in terms of citations across all articles published in the International Journal of Forecasting.

Almost 20 years have passed since the previous instalment in the M competition series, the M3. Its 3,003 series have since then been used numerous times and occupy the empirical design sections of many published forecasting papers. However, one would argue that forecasting approaches now tend to overfit the M3 data while hyperparameters are, in some instances, ad-hoc selected to achieve the desired accuracy of the (known) post-sample data.

The M4 competition builds on the success of the previous ones, bringing a challenging task of forecasting 100,000 series. The main innovations of the M4 competition aim to address criticisms regarding the design of previous competitions: more data frequencies are considered, prediction intervals are also evaluated, and statistically robust error measures are included. In

addition, the M4 competition sets to be more transparent than any other forecasting competition before, with most of the participants submitting the code of their methods and the organizing team making publicly available the procedure and code used to evaluate the submitted forecasts and perform the respective analyses, Moreover, and the organizers replicating the majority of the submitted the results by the submitted methods.

Regardless of the innovations of the M4 competition, many would still argue that it is not fully representative of reality. We ask ourselves: “how can a forecasting competition be representative of reality?” Maybe one that would feature all real data that need to be forecasted (together with quantitative or qualitative information that might be of value) and where participants would ask to forecast in real time? In our opinion, there will never be a “perfect” forecasting competition. In fact, a “perfect” forecasting challenge is what forecasting practitioners face every day. Still, forecasting competitions, similar to laboratory experiments, can enhance our understanding on what affects the accuracy of forecasting methods and become catalysts in the development of unique approaches that advance the art and science of forecasting.

In this issue…

This special issue presents the M4 competition. The editorial is followed by a brief history of forecasting competitions (Hyndman, 2019) that includes interesting anecdotes on the background of the early M competitions, as well as a summary of the major statistical forecasting competitions and their findings to date. Then, it provides an overview of the state of forecasting in social settings (Makridakis, Hyndman and Petropoulos, 2019) so that readers who are less familiar with the forecasting literature will be able to better understand the latest competition and put its results into context.

These three introductory articles are followed by the main course: The predictions/hypotheses of the findings of the M4 competition before such findings were available (Spiliotis, Makridakis and Assimakopoulos, 2019), followed by the main paper with the actual results of the M4 competition (Makridakis, Spiliotis and Assimakopoulos, 2019). This paper provides a detailed analysis of the results of the competition, rankings of all submitted methods for point forecasts and prediction intervals, analysis of the statistical differences between methods, degree of replicability/reproducibility of the submissions and disaggregation of the results for different categories and frequencies of data. The article concludes with what we have learned from this large empirical exercise and what the next competition should aim to investigate.

After the presentation of the main results, the reader is exposed to the methodologies of the submissions with the top performance, either overall or for specific categories. In the next section, we overview the invited submissions and provide details on the criteria for inclusion of a method in this special issue.

We then turn our attention to the *why* of the results: six discussion papers, with authors from both academia and industry, crutinize the results of the competition to provide useful insights, interpretations and criticisms. Following the discussion papers, ten invited commentaries serve the basis for a debate on the design and the results of the M4 competition, and competitions in

general. We further discuss the contributors of the discussion papers and commentaries in the penultimate section of this editorial. The final, concluding paper of this special issue is a rebuttal to the discussion papers and commentaries by the organizers of the M4 competition.

The winning submissions

The inclusion process of the methods and approaches that were submitted to the M4 competition and appear in this issue was done according to the following criteria. First, we invited the best pool of methods whose performance did not statistically differ in terms of Multiple Comparisons from the Best (MCB; see Koning et al., 2005). Then, we invited methods and approaches that scored a place within the top three in terms of the prediction intervals, any of the data categories (macro, micro, demography, industry, etc.) or any of the frequencies (yearly, quarterly, monthly, etc.). We excluded methods submitted by participants directly associated with the Forecasting & Strategy Unit of the National Technical University of Athens (FSU-NTUA) that co-organized the M4 competition. The above process provided ten methods that are presented in Table 1, along with the reason for their inclusion.

Table 1. Invited methods and approaches in the M4 competition special issue

Author(s)	**Affiliation**	**Reason for inclusion**
Smyl	Uber	Not statistically different, based on the MCB
Montero-Manso et al.	University of A Coruña & Monash University	
Pawlikowski et al.	ProLogistica Soft	
Jaganathan & Prakash	Individual	
Fiorucci & Louzada	University of Brasilia & University of São Paulo	
Petropoulos & Svetunkov	University of Bath & Lancaster University	2nd for weekly data 3rd for quarterly data
Shaub	Harvard Extension School	3rd for yearly, after excluding participants from FSU-NTUA
Doornik et al.	University of Oxford	1st for hourly 3rd for prediction intervals
Ingel et al.	University of Tartu	2nd for daily
Darin & Stellwagen	Business Forecast Systems (Forecast Pro)	1st for weekly

A short description of each of these submissions is provided below.

Smyl (2019) methodologically combines traditional statistical methods (Holt-Winter exponential smoothing) with NN. This approach also utilizes cross-learning to optimize the parameters both locally and globally (across series) through a hierarchical approach. Montero-Manso et al. (2019) propose a weighted combination approach where the weights for combining the available methods for each series are derived through meta-learning based on a large number of time series features. One common element of these two approaches is that univariate forecasting is improved using cross-sectional information.

The methods by Pawlikowski et al. (2019), Jaganathan and Prakash (2019) and Fiorucci and Louzada (2019) are combinations based on past forecasting performance. Pawlikowski et al. (2019) and Jaganathan and Prakash (2019) suggest the use of different pools of models per frequency, while Fiorucci and Louzada (2019) combine the forecasts from three models: Theta, exponential smoothing and ARIMA. Pawlikowski et al. (2019) and Fiorucci and Louzada (2019) determine the combination weights based on cross-validation. Jaganathan and Prakash's (2019) combination approach has the form of pooling (including/excluding models based on their performance) but they also propose for some frequencies an evidence-based ensemble where forecasts from a predefined pool of models (different for each frequency) are combined using the median operator. Petropoulos and Svetunkov (2019) perform a median combination of the point forecasts and the prediction intervals too; however, their approach consists of just four models (the same models for all frequencies). In a similar fashion, Shaub (2019) forecasted the yearly subset of the M4 data using the simple arithmetic mean of the forecasts of three models: ARIMA, TBATS and Theta.

The first place in the hourly frequency and third in the predictions intervals was achieved by Doornik et al. (2019), who developed a forecasting method that calibrates the average of a simple yet adaptive autoregressive model (Rho) and a damped trend estimation of the growth rate (Delta). The Tartu team's second place on the daily frequency (Ingel et al., 2019) was based on a combination of five statistical models as well as a simple observation that some of the daily data are highly correlated. Finally, Darin and Stellwagen (2019) achieved the first place in the weekly subset by using their proprietary software, Forecast Pro, to obtain baseline forecasts and appropriately determining a performance indicator that matched the error measure used in the competition as well as appropriately reviewing and revising cases where the baseline forecasts could be inadequate.

Even if, due to limited space, we only present here 10 of the 49 submissions of the M4 competition, we believe that all individuals and teams that participated in the competition are winners for two reasons. First, the task involved in forecasting a very large number of series can be compared to finishing a Marathon. Second, as uncertainty is inherent in any forecasting competition, participants in risk their reputations if their method does not do well. However, those interested in advancing the field should not be scared in making their predictions public and be evaluated against those of others. The same goes for assessing uncertainty, introduced for the first time in an M Competition, aiming at estimating 95% Prediction Intervals (PIs) and being evaluated against others knowing the possibility of underestimating reality.

We consider all entries to be part of the real legacy of the M4 competition. The design of the methods and approaches related to these submissions contained a large number of innovative ideas and most importantly a tremendous amount of work. We can confidently predict that the forecasting community will make use of these innovations and will come up, in the not too distant future, with new approaches, combining structural features and even forecasts of the submitting methods. For instance, the first author found that a simple, median combination[1] of the point forecasts from the top-6 submissions performs statistically better in terms of MCB (based on mean ranks) than any of the entries in the M4 competition and is on par with the top six submissions in terms of accuracy measures used in the M4 competition.

Discussion papers and commentaries

The special issue continues with the following insightful discussion articles:

- Barker (2019) from Microsoft discusses the performance of machine learning (ML) models and the differences between structured and unstructured models.
- Januschowski et al. (2019) from Amazon argue that there does not exist a major methodological distinction between ML and statistical models; this yields the need that these two communities (statistics and ML) need to communicate more effectively.
- Fry and Brundage (2019) from Google provide a list of five real forecasting challenges as suggestions for future forecasting competitions.
- Gilliland (2019) from SAS discusses the added value (and added cost) of ML methods in the forecasting process.
- Lichtendahl and Winkler (2019) provide insights on what makes some forecast combinations more successful than others.
- Grushka-Cockayne and Jose (2019) take another look at the evaluation of the submitted prediction intervals and suggest that intervals combination improves the performance of individual submissions.

Finally, the special issue includes ten invited commentaries from leading academics and researchers (in alphabetical order): Amir Atiya, Gianluca Bontempi, Robert Fildes, Paul Goodwin, Tao Hong, Ioannis Katakis et al., Stephan Kolassa, Konstantinos Nikolopoulos et al., Dilek Onkal and Keith Ord.

Acknowledgments

We would like to thank everyone that helped with the production of this special issue: the former Editor-in-Chief, Rob Hyndman, the current Interim Editor-in-Chief, Esther Ruiz, the contributing authors and finally the Associate Editor and reviewers who had the difficult task of reviewing the papers included in this special issue.

References

Barker, J. (2019). Machine Learning in M4: What Makes a Good Model? *International Journal of Forecasting*, *35*(4), XXX-XXX.

[1] Similar to the simple combination of univariate models, proposed by Petropoulos and Svetunkov (this issue)

Darin, S., & Stellwagen, E. (2019). Forecasting the M4 Competition Weekly Data: Forecast Pro's Winning Approach. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Doornik, J., Castle, J., & Hendry, D. (2019). Card forecasts for M4. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Fiorucci, J. A., & Louzada, F. (2019). GROEC: Combination method via Generalized Rolling Origin Evaluation. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Fry, C., & Brundage, M. (2019). The M4 Forecasting Competition - A Practitioner's View. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Gilliland, M. (2019). The Value Added by Machine Learning Approaches in Forecasting. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Grushka-Cockayne, Y., & Jose, V. R. R. (2019). The M4 Forecasting Competition Prediction Intervals. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Hyndman, R. (2019). A brief history of forecasting competitions. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Ingel, A., Shahroudi, N., Kangsepp, M., Tattar, A., Komisarenko, V., & Kull, M. (2019). Correlated daily time series and forecasting in the M4 competition. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Jaganathan, S., & Prakash, P. (2019). Combination based forecasting method: M4 competition. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Januschowski, T., Gasthaus, J., Flunkert, V., Wang, B., Bohlke-Schneider, M., Salinas, D., & Callot, L. (2019). Criteria for Classifying Forecasting Methods. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409.

Lichtendahl Jr., K. C., & Winkler, R. (2019). Why Do Some Combinations Perform Better Than Others? *International Journal of Forecasting*, *35*(4), XXX-XXX.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Hyndman, R., & Petropoulos, F. (2019). Forecasting in social settings: the state of the art. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Makridakis., S., Spiliotis, E., & Assimakopoulos V. (2019). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Montero-Manso, P., Talagala, T., Hyndman, R., & Athanasopoulos G. (2019). FFORMA: Feature-based Forecast Model Averaging. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Pawlikowski, M., Chorowska, A., & Yanchuk, O. (2019). Weighted Ensemble of Statistical Models. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Petropoulos, F., & Svetunkov, I. (2019). A Simple Combination of Univariate Models. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Shaub, D. (2019). Fast and Accurate Yearly Time Series Forecasting with Forecast Combination. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Smyl, S. (2019). Exponential Smoothing and Recurrent Neural Network Hybrid Model. *International Journal of Forecasting*, *35*(4), XXX-XXX.

Spiliotis, E., Makridakis., S., & Assimakopoulos V. (2019). Predicting/Hypothesizing the findings of the M4 Competition. *International Journal of Forecasting*, *35*(4), XXX-XXX.